图像灰度密度分布计算模型及肺结节良恶性分类 *

Vanbang Le[1]，朱 煜[1†]，郑兵兵[1]，杨达伟[2]，任晓东[1]，Thiminhchinh Ngo[3]

(1. 华东理工大学 信息科学与工程学院，上海 200237; 2. 复旦大学附属中山医院，上海 200032; 河内高科技研究中心，越南 河内)

摘 要：计算机辅助肺癌诊断对于肺癌的早期发现及提早治疗具有重要意义。提出一种基于密度分布的特征评估算法，同时引入模式识别模型来评估该方法的效率。首先，从肺部肿瘤图像中随机提取像素块集，通过 K-均值聚类算法将其分为 10 类，根据 CT 图像中肺结节像素值和聚类中心的关系，提取出 10 维特征向量，利用随机森林分类器进行模型训练，进而判断肺结节良恶性水平。通过 CT 图像公开数据集 LIDC-IDRI 实验表明分类平均精度达到 0.9008。实验结果对比分析表明，提出的特征表达方法具有更优的分类效果和更高的鲁棒性。

关键词：肺结节分类；密度分布特征；K-均值

中图分类号：TP391.41　　doi: 10.19734/j.issn.1001-3695.2018.05.0505

Pulmonary nodule image grey density distribution feature extraction algorithm and adenocarcinoma benign/malignant classification

Vanbang Le[1], Zhu Yu[1†], Zheng Bingbing[1], Yang Dawei[2], Ren Xiaodong[1], Thiminhchinh Ngo[3]

(1. *School of Information Science & Engineering*, *East China University of Science & Technology*, *Shanghai* 200237, *China*; 2. *Zhongshan Hospital*, *Fudan University*, *Shanghai* 200032, *China*; 3. *Hitech Telecommunication Center*, *Hanoi Vietnam*)

Abstract: Aimed-at lung nodule Benign/Malignant classification, an effective grey scale density distribution feature extraction algorithm which was combined with pattern recognition models to evaluate the classification system was proposed. The proposed feature extraction algorithm first collected a large number of blocks from lung tumor images and determined the distance matrix by calculating the relationships among the image blocks. Then, K-means clustering methods was used to classify the current image blocks and obtained 10 cluster centers. After that, calculated the distribution density features by mapping CT value of nodule image pixels with the 10 cluster centers and extracted a 10-dimensional feature vector. Finally, the extracted feature vectors were divided into training and testing set to identify lung adenocarcinomas risk levels by Random Forest classification model. The classification framework was evaluated in LIDC-IDRI dataset, the average accuracy reached to 0.9008. The proposed method outperforms the most recent techniques, and the experimental results show great robustness of the proposed method for different lung CT image datasets.

Key words: lung nodule classification; density distribution feature; K-means

0 引言

根据世界各大癌症研究中心和卫生组织的调查显示，肺癌已经成为全世界致死率最高的第一大癌症。目前胸腔扫描图像技术的应用范围越来越广泛。通过分析 CT（computed tomography）图像的特征发现早期肺结节并且及时给出正确的诊断、治疗，从而提高病患者的生存率，因此通过数字图像处理技术对肺部 CT 图像进行分析已经成为热点研究方向。肺部 CT 图像计算机辅助诊断系统中主要包含肺结节检测，分割，分类等研究项目。其中提升微小肺腺结节（病灶直径<30mm）的诊断和识别水平能显著的提高肺癌的诊断准确率以便为临床医生提供更加准确的诊断建议，所以其一直是图像处理领域的重要课题[1]

辅助诊断系统中肺结节分割一直为最重要的步骤之一，对

收稿日期：2018-05-09；**修回日期：**2018-07-09　**基金项目：**国家自然基金青年基金资助项目（81500078）；复旦大学附属中山医院临床研究专项基金资助项目（2016ZSLC05，2016ZSCX02）

作者简介：Vanbang Le（1988-），男，博士研究生，主要研究方向为医学图像处理与模式识别；朱煜（1973-），女（通信作者），教授，博士，主要研究方向为图像与视频处理、机器学习（zhuyu@ecust.edu.cn）；杨达伟（1985-），男，主治医师，主要研究方向为呼吸内科肺部恶性肿瘤疾病研究及基于人工智能的临床肺小结节肺癌早期诊断；郑兵兵（1992-），男，博士研究生，主要研究方向为图像与深度学习；任晓东（1992-），男，硕士研究生，主要研究方向为计算机视觉与机器学习；Thiminhchinh Ngo（1988-），女，研究员，主要研究方向为计算机科学、信号处理.

肺结节良/恶性分类影响很大。常见的肺结节的分割方法为灰度阈值法、GRAPH-CUTS、水平集、深度学习等。而主要的算法验证数据为 LIDC-IDRI[2]、ELCAP[3]、NLST[4]等。借助分割后的结节图片能够在一定程度上评估结节的后续增长趋势及良恶性病变水平[5]。

肺结节良/恶性分类旨在给医生提供科学、可靠的辅助分类结果，使诊断过程更加精准并且有效的降低医生的阅片工作量。肺结节分类的基础为图像特征提取方法，通过图像特征并与分类器结合进行训练和测试。而常见的分类器主要包含 SVM[6]、KNN[7] 、随机森林[8]等。临床医学中从 CT 值分布角度可将肺结节分为磨玻璃型、半实质型及实质型的，而从危险程度来看可分为良性和恶性肺结节。Han 等人[9]以 LIDC 数据库为研究对象，通过提取肺结节的 2D/3D 纹理（Harralick 纹理特征）及几何特征（圆度、外接矩形充实度等）将肺结节分为良性/恶性两类。实验结果的最大 ROC 指数为 92.7%。Dhara[10]根据肺结节的几何和 Harralick 纹理特征将 LIDC-IDRI 数据的样本集分为良性和恶性两类，其最优 AUC（Area Under Curve）值达到了 0.9505。康奈尔大学的 Reeves[11]使用 46 维空间特征对 PLIB（public lung image database）和 NLST（national lung screening trial database）实现肺结节良恶性分类。实验表明在参数最优的前提下，其分类准确率达到了 70%。梅奥医疗中心生理与生物医学团队在研究成果中介绍了 CANARY（computer aided nodule assessment and risk yield）系统[12]，其对 NLST 进行密度聚类分析，并对病患经过 5 年的跟踪研究，提出了计算机辅助肺结节分类与风险预测结论。Maldonado[12]提出一种肺结节图像密度分布计算方法用于 CANARY 的分类模块中，该特征描述肺结节的 HU（hounsfield unit）值分布情况，非常有借鉴意义。太原理工大学的裴博[15]使用基于双向隶属度函数的模糊支持向量机，综合考虑肺结节的灰度、纹理及形状特征，实现了 83%的识别准确率和 10%的误诊率。

为提高肺结节良/恶性的分类性能，本文中提出一种基于图像子块集的肺结节图像灰度密度分布特征提取模型。首先从肺结节图像集中获取子块集，计算该数据集的自相关矩阵并使用无监督聚类算法对自相关矩阵进行聚类。从而获得图像子块对应的标签，然后通过寻找目标测试像素最匹配的子块计算测试图像每一像素的标签。最后统计、生成肺结节图像的灰度密度分布特征，并结合随机森林分类器对数据集分类。

1　材料与方法

1.1　实验材料

LIDC-IDRI 肺部 CT 公开数据库（The lung image database consortium and image database resource initiative）为目前较大、常用的肺 CT 公开数据库。LIDC-IDRI 数据库从 The Cancer Imaging Archive (TCIA) 官网下载，肺结节的边缘坐标及特征可以从附带的*.XML 文件中提取。LIDC-IDRI 中，肺部 CT 图像尺寸均为 512×512（单位为 HU 值），其重要的参数为：Slice Thickness 表示切片厚度（单位为毫米）；Pixel Spacing 表示像素中心间的物理间距（单位为毫米）。肺结节的精确边缘坐标及其特征标注一般由 4 位放射科医生实现，标注结果很显然是存在一定的差异的。从 XML 文件获取肺结节区域时，被选择的目标标注为最大面积的区域，即

$$nodule = \arg\max\left(mark^{(i)} \mid i = 1\sim4\right) \tag{1}$$

本文中从 LIDC-IDRI 大规模的肺结节样本数量选取出一部分作为研究对象，选取对象的共同点是均为小型肺结节，最大长径均小于 20mm，共 1285 肺结节样本。研究对象数据中，像素间距及片间距分布分别从 0.5 至 0.8mm 与 0.6mm 至 5.0mm，长径范围为[2.79mm, 15.77mm]。肺结节最重要的评估参数——危险程度被分为 5 个等级（rank 1~5），实验中其样本数量分别为 147/390/387/250/119，共 1 285 个肺结节。使用 Han[9]的良恶性规划方案，形成 3 种良恶性先验定义方案，分别为 Configuration 1,2,3 （CF 1, 2, 3）。其中，CF1 的良性与恶性结节分别由 rank 1, 2 和 rank 4, 5 的样本组成；CF2 的良性样本为 rank 1, 2, 3 组成，恶性样本为 rank 4 和 5；CF3 则将 rank 3 的肺结节定义为恶性，即 rank 1, 2 的样本为良性，rank 3, 4, 5 的样本为恶性。LIDC-IDRI 数据集样本详细信息统计如表 1 所示。

表 1　LIDC-IDRI 部分样本及不同数据子集基本信息

Table 1　LIDC-IDRI partial samples and basic information of different subset of data

Sub-sets	Cfs.	Benign		Malignant	
		Ranks	样本	Ranks	样本
Sub-set 1	CF1	‘1’~‘2’	475	‘4’~‘5’	180
(LS 1)	CF2	‘1’~‘3’	802	‘4’~‘5’	180
(3-10)mm	CF3	‘1’~‘2’	475	‘3’~‘5’	507
Sub-set 2	CF1	‘1’~‘2’	54	‘4’~‘5’	189
(LS 2)	CF2	‘1’~‘3’	114	‘4’~‘5’	189
(10-20)mm	CF3	‘1’~‘2’	54	‘3’~‘5’	249
	CF1	‘1’~‘2’	529	‘4’~‘5’	369
All nodules	CF2	‘1’~‘3’	916	‘4’~‘5’	369
	CF3	‘1’~‘2’	529	‘3’~‘5’	796

针对不同尺寸下的分类性能分析，本文将肺结节进一步分为两个数据子集 Sub-set 1 和 2（LIDC-IDRI Sub-set, LS），其中 LS1 中的肺结节尺寸范围从~3mm 至~10mm 而 LS2 肺结节的从~11mm 至~20mm。通过对统计 LS 和 LS2 分类性能的差异及详细的分析，本文总结了图像特征对识别不同尺寸肺结节的影响。

1.2　肺结节图像灰度密度分布特征

肺部 CT 影像中，可疑区域的灰度级分布影响到肺结节的定位和分类。因此灰度密度分布是肺结节图像危险程度重要的判决指标之一。图像的灰度密度分布指的是图像中像素值与周围邻近点之间的关系，其表征着图像任意局部区域灰度值出现的强度及其幅度。图像中密集出现高灰度值的区域为高密度区域，而高灰度值像素较稀疏的则为低密度区域。

1.2.1 图像单元库获取方法及其自相关矩阵

肺部 CT 图像单元数据库（Image block database, IBD）是本文提出的肺结节图像特征提取过程中作为最重要的环节。这些图像单元是从肺结节图像中随机提取出来，其数量和多样性将决定特征的表征精度。为确保IBD的多样性及使其平衡，在不同的肺结节数据库（LIDC-IDRI 和 ZSDB）的各类别中挑选出平衡数量的图像单元。对图像单元大小而言，因为目标处理的肺结节图像大小范围为 3mm~30mm，图像单元不能太大也不能太小。如果太小，处理之后更接近于点处理的结果从而会引入噪声，太大则对较小的肺结节带来较大误差，因此一般采用 5×5、7×7 或 9×9 的图像单元大小。单元大小会影响到目标计算图像的平滑效果，较大的单元适用于大型肺结节，而对小型肺结节不能使用太大（尺寸）的图像单元集。在每一个肺结节图像随机提取图像单元构造出各类数量均衡的 IBD。

为对IBD进行聚类，先计算两两图像单元之间的距离并生成 IBD 的距离矩阵 $h(i,j)$。设 IBD 中的图像单元数量为 N，则 $h(i,j)$为大小为 N×N 的对称方阵。$h(i,j)$中每一行（或列）为某个单元与其他单元的距离向量。向量之间通常的距离计算方法包含 Euclid (EU)、Canberra (CA)、Chebyshev (CH)、Braycurtis (BR)等多种模式，为保证距离值之间具有最大的区分度本文中使用 Canberra 距离来计算向量之间的距离。任意两个等长的向量 p 和 q 的 Canberra, Euclid, Chebyshev 及 Braycurtis 的距离计算公式如下：

$$d_{CA}(p,q)=\sum_{i=1}^{n}\frac{|p_i-q_i|}{|p_i|+|q_i|} \tag{2}$$

$$d_{EU}(p,q)=\sqrt{\sum_{i=1}^{n}(p_i+q_i)^2} \tag{3}$$

$$d_{CH}(p,q)=\max_{i}|p_i-q_i| \tag{4}$$

$$d_{BR}(p,q)=\frac{\sum_{i=1}^{n}|p_i-q_i|}{\sum_{i=1}^{n}|p_i+q_i|} \tag{5}$$

其中 n 为向量的长度。

本文中通过无监督聚类方法对距离矩阵进行归类，聚类中心数量为 10。实验表 K-Means[14] 聚类方法返回结果的 Silhouette[13]（SIL）是最优的。将聚类结果映射至 IBD 并找出对应图像单元的标签，这样每一个单元都被标上记号并生成被标记的图像单元数据库（Marked Image Block Database, MIBD）。此时，距离矩阵的每一个行（或列）的聚类结果同时也是距离向量所对应的图像单元的灰度密度分布等级。IBD 聚类过程中，图像单元、距离矩阵及聚类统计结果示意图如图 1 所示。

图 1 中，图 1（A）为 IDB 的可视化图像，其中 1600 个图像单元排成[(40×7)×(40×7)]的矩阵，图中蓝色和深橙色代表图像单元最小与最大的灰度值。图 1（B）的 1, 2, 3, 4 分别为不同向量间距离计算方法（BR, CA, CH, EU）的自相关矩阵（左边）及其对应的聚类标签（右边），而对应的直方图（bins=256）如图 1（C）所示。通过聚类距离矩阵计算出对应图像单元的标签（K=10）。图像单元集 IDB 聚类标签统计如图 1(D)所示，从图中的分布曲线可以看出由 CA 和 BR 计算得到距离矩阵的聚类统计分布较平滑。为定量分析聚类结果之间的优异与否，本文使用 Silhouette（SIL）参数计算聚类的相对精度。SIL 参数由类间距离及类内的紧凑度构成，分布范围为[-1，1]。SIL 越大表明性能越好（最好时为 1）。CA 的对应 SIL 为最大 0.4310，因此本文中使用 CA 距离计算方法对图像单元集进行自相关分析。本文中 IDB 的标签已按类别中心的值从小至大进行排序。

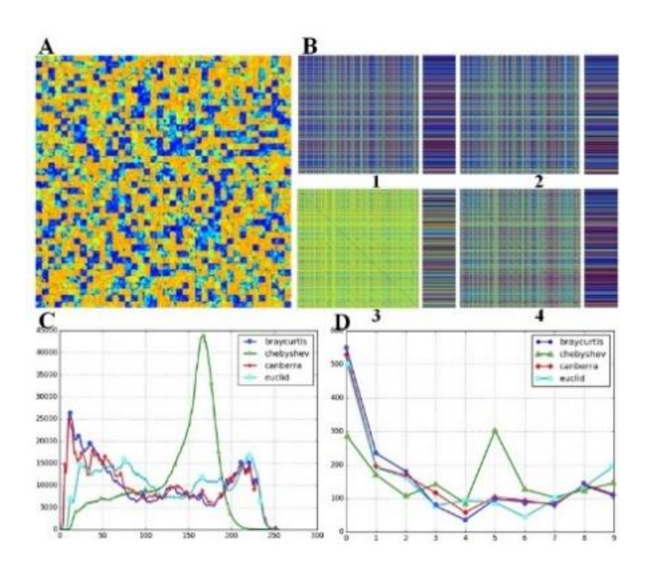


图 1　图像单元集图像及其自相关矩阵与聚类统计结果

Fig.1　Image unit set image and its autocorrelation matrix and clustering statistical results

1.2.2 基于图像单元灰度密度分布特征获取

IBD 进行聚类后，对肺结节图像遍历计算每一个非背景像素的灰度密度分布等级。在此过程中，以目标像素为中心提取遍历窗口 $I_{test}(x,y)$，窗口大小等于 IBD 中图像单元的大小。通过欧式距离的计算，搜寻 IBD 中与之匹配的单元（距离最小的），记为 $I_{matched}(x,y)$，见式(6)

$$I_{matched}(x,y)=\arg\min_{i\in[1,N]}(\|IBD_i,I_{test}(x,y)\|) \tag{6}$$

此时，$I_{matched}(x,y)$ 在聚类结果中的标签为 $I_{test}(x,y)$ 的灰度密度分布等级，即

$$Level(I_{test}(x,y))=Label(I_{matched}(x,y)) \tag{7}$$

依次计算肺结节图像中全部非零像素的密度分布等级最终得到肺结节的 CT 值密度分布图像。该图像的有效值数量为 10（1~10），在这过程中将灰度密度分布等级代替了像素原有的值。本文将密度分布特征作为肺结节的识别特征，用于机器学习训练的标准特征向量集。特征提取过程示意图如图 2 所示。

图 2 中，(a)为肺结节图像，(b)为肺结节图像的灰度密度分布图像，(c)为饼图表示的特征向量。通过密度分布图像可以看出肺结节内的 HU 值高低分布情况有助于定量估计实质部

分的位置及大小从而提高临床医生对肺结节分类的准确性。

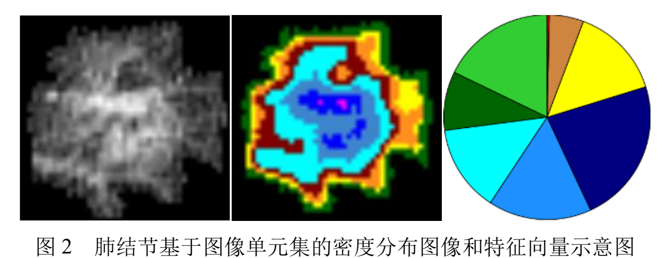


图 2 肺结节基于图像单元集的密度分布图像和特征向量示意图

Fig.2 Lung nodule based on image unit set density distribution image and eigenvector schematic map

1.3 随机森林分类器及模式识别评价

本文中使用随机森林（Random forest）方法作为分类器，该分类器是包含多个决策树的分类器。分类评价参数为敏感度（Sensitivity），特异性（Specificity），ROC（Receiver operating characteristic）曲线及识别精度（Accuracy）。其中，敏感度（Sensitivity）或真阳性率（True positive rate, TPR）为描述分类性能的“阳性”样本正确判别率，数学模型为

$$Sensivity = \frac{TP}{TP+FN} \tag{8}$$

特异性（specificity）或 真阴性率（true negative rate, TNR）统计分类性能的“阴性”样本正确判别率，数学公式为：

$$Specificity = \frac{TN}{TN+FP} \tag{9}$$

其中：真阳性：（true positive, TP） = 正确判断为真；假阳性：（false positive, FP） = 错误的判断为真；真阴性：（true negative, TN） = 正确判断为假；假阴性：（false negative, FN） = 错误的判断为假。

ROC 曲线通过图像方式对二分类模型进行评价，由真阳性率（TPR）及假阳性率（FPR）构成。 ROC 曲线下的面积 AUC（area under the curve，AUC）被用来量化 ROC 曲线。AUC 值越大说明分类器性能越优。识别精度 Accuracy 的计算公式为：

$$Accuracy = \frac{\sum TP + \sum TN}{AllSamples} \tag{10}$$

2 实验结果与分析

本节展示验证数据库的分类性能的实验效果及其分析结果，分类模型实验配置具体如下：分类器为随机森林（RF），estimator = 100；对于 LIDC-IDRI 的训练和测试样本比例为 50:50（%）；每一个子集分别做 100 次实验并计算平均的性能评价参数值。实验平台配置：编程语言：Python 3.0-Windows 10；硬件信息：Processor Intel(R) Core(TM) i7-6700HQ 2.60GHz (8CPUs); GPU Geforce 960; RAM 8Gb。实验过程中，提出的系统运算效率较高并满足实时性处理要求，单样本分类的平均运行时间为 35±0.5（ms）。

2.1 肺结节图像特征提取与分析

LIDC-IDRI 数据库中的样本聚类结果展示图如图 3 所示。

从图 3 中可以看出肺结节的密度图中各个类别一般以环形围绕着中心点，最外围为最小值（k=1），因此在特征向量中第一分量均为非零的。对于 LIDC-IDRI 数据库特征向量中低等级密度占据的比例从 rank 1 至 rank 5 稳定递减而高等级的密度比例则为递增的。LIDC-IDRI 中高分布等级比例的排序为 rank 1 < rank 2 < rank 3 < rank 4 < rank 5。LIDC-IDRI 肺结节的灰度密度分布特征具有可靠的统计意义，各类特征向量之间差异很明显且稳定，非常有助于提高肺结节良恶性的分类性能。LIDC-IDRI 样本的灰度密度分布特征均值如图 4 所示。

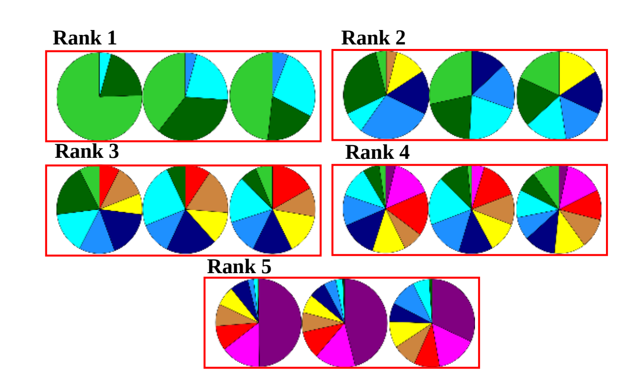


图 3 基于图像子块集灰度密度分布特征示意图

Fig.3 Gray scale density distribution based on image subblock

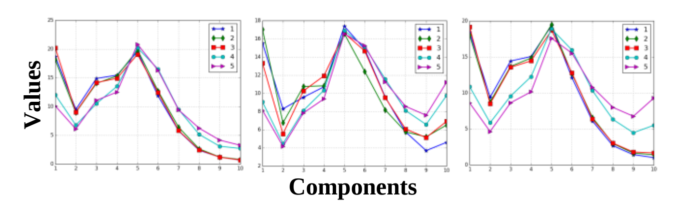


图 4 密度分布平均特征向量（左：LS1；中间：LS2；右：LS1+LS2）

Fig.4 Average eigenvector of density distribution (left: LS1; middle: LS2; right: LS1+LS2)

对于 LIDC-IDRI 数据库，小型和大型肺结节的良恶性特征分布虽然具有一定的区分度但差异并不明显，rank 1 至 rank 5 平均曲线的趋势较类似，尤其是直径大于 10mm 的 LS2。对整体的 LS1+LS2 而言，rank 1 至 rank 5 平均向量前四个分量（密度值小于-663HU）之和分别为 51.3%，55.24%，55.63%，38.52%和 32.01%而最后 4 个分量（密度值大于-282HU）的总和依次为 11.19%，12.57%，12.85%，26.61%和 34.83%。其中后四个分量（代表实质性区域比例）的排序均为 rank 1<rank 2< rank 3< rank 4< rank 5。从而可以看出 rank 1~5 的特征向量分布趋势为低等级分量比例逐渐降低而高等级密度分量则相反。

综上所述，临床应用中通过观察肺结节的密度分布图以及统计密度特征可以更直观地表达病灶的结构，有助于提高诊断效率以及分类精度。

2.2 肺结节图像的分类

实验结果表明整体样本集上（LS1+LS2）分类性能最好的是 CF1，而当 rank 3 的肺结节被分配至良或恶性类别（CF2 与 CF3）时评价指数低一些。分类评价参数的 AUC，敏感度和特异性统计如表 2 所示。

表 2　LIDC-IDRI 分类性能评价参数统计

Table 2　Evaluation parameter statistics of LIDC-IDRI classification performance evaluation parameter statistics

LS1+LS2 (所有样本)			
Cfs.	AUC	敏感度	特异性
CF1	0.9681±0.0055	0.9333±0.0255	0.8786±0.0250
CF2	0.9405±0.0065	0.8742±0.0317	0.8800±0.0137
CF3	0.8070±0.0129	0.7239±0.0297	0.7296±0.0476
LS1-较小的结节			
CF1	0.9820±0.0043	0.9114±0.0324	0.9475±0.0132
CF2	0.9702±0.0045	0.8661±0.0413	0.9436±0.0108
CF3	0.7681±0.0148	0.5958±0.0402	0.8047±0.0573
LS2-较大的结节			
CF1	0.9273±0.0196	0.9347±0.0416	0.5894±0.1452
CF2	0.8203±0.0263	0.8897±0.0494	0.5934±0.0671
CF3	0.8941±0.0219	0.9603±0.0238	0.3839±0.1497

Cfs.：样本规划方案（Configurations）

由表 2 看出对于 LS1，性能评价排序为 CF1>CF2>CF3 而对于 LS2 为 CF1>CF3>CF2。从而可以判断 LIDC-IDRI 中的 rank 3 小型肺结节较倾向于良性的，而大型的则倾向于恶性的。由于 LS1 的样本较大（LIDC-IDRI 中多数肺结节均小于 10mm），故整体数据上 rank 3 的肺结节更具有良性肺结节的特征。

LIDC-IDRI 测试样本的识别精度对于 CF1，CF2 和 CF3 分别为 0.9008，0.8782 与 0.7258。其中阳性/阴性预测交叉矩阵如图 5 所示。其中，Bn 表示良性（Benign）；Ml 表示恶性（Malignant）。左至右分别为 CF1，CF2 和 CF3)。)。由于 rank 3 肺结节的干扰，CF1 的分类性能较稳定于 CF2 和 CF3。

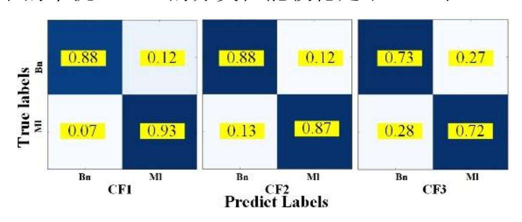


图 5　LIDC-IDRI 的平均交叉检验矩阵

Fig.5　mean cross test matrix of LIDC-IDRI

如上述所说，rank 3 的样本对 LIDC-IDRI 良恶性分类性能影响很大，使用的数据中 rank 3 肺结节数量越多在理论上分类效果越不稳定。对于实验数据结构而言，相较于 Han 等人[9]（ 172 样本）和 Dhara [10]（349 样本），本文使用的 rank 3 样本数量为 387，数据集给分类模型带来的困难较大。虽然如此，利用本文所提出的灰度密度分布特征，测试数据时仍然获得非常可观的评价指标。其中 CF1，CF2 和 CF3 的平均 AUC 分别为 0.9681，0.9405 和 0.8070。本文提出灰度密度分布特征与几何和纹理特征的 ROC 对比如表 3 所示。由表 3 中可以看出虽然 rank 3 样本数量较大但是本文提出的分类模型在三种样本先验规划下性能均略高于文献[9]和[10]提出的基于几何和纹理特征的分类模型。因此表明所提出的灰度密度分布特征对肺结节良恶性分类的有效性，同时也证明了所提出特征的分类性能优于图像 2D/3D 几何和纹理特征。

表 3 分类性能与对比

Table 3　Comparison of classification performance

Configurations	CF1	CF2	CF3
本文特征	0.9681	0.9405	0.8070
几何和纹理特征 *	0.8784	0.8108	0.7210
文献[10]	0.9505	0.8822	0.8488
文献[9]	0.9450	0.8703	0.8315

*本文数据使用 Dhara[10]提出几何和纹理特征时的 ROC.

3　结束语

本文中提出一种基于图像子块集的肺结节灰度密度分布特征计算方法。首先从肺结节图像集随机挑选得到一致大小的若干图像单元构成单元集。然后计算该数据集的自相关矩阵并将距离矩阵聚为 10 类，获得对应图像单元的聚类标签。最后通过遍历肺结节图像，每一个像素与周围邻近点构成的窗口与图像单元集对照，寻找最匹配的单元，该单元的标签则为测试像素的灰度密度分布等级。最终统计肺结节图像的密度分布图获取其特征。实验结果与对比分析表明，基于密度分布的特征评估算法具有能有效的对肺结节良恶性等级进行分类的能力。本文的研究结果为肺结节临床辅助诊断提供了新的方法，同时也对中国或亚洲地区的肺癌早期诊断系统的发展有参考价值。

参考文献：

[1] 张满. 孤立性肺结节良恶性预测模型的建立 [D]. 广州：南方医科大学, 2016. (Zhang Man. Establishment of a mathematic model for predicting malignancy in solitary pulmonary nodules [D]. Guangzhou: Southern Medical University, 2016.)

[2] Armato S G, McLennan G, *et al.* The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans [J]. Medical physics, 2011, 38 (2): 915-931.

[3] Welch H G, Woloshin S, Schwartz L M, *et al.* Overstating the evidence for lung cancer screening: the international early lung cancer action program (I-ELCAP) study [J]. Archives of Internal Medicine, 2007, 167 (21): 2289-2295.

[4] National Lung Screening Trial Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening [J]. New England Journal of Medicine, 2011, 365 (5): 395-409.

[5] 邢谦谦. 不规则形态肺结节的分割及毛刺检测研究 [D]. 广州：南方医科大学, 2015. (Xing Qianqian. Research on irregular lung nodule automatic segmentation and spiculation detection [D]. Guangzhou: Southern Medical University, 2015.)

[6] Tan Yongqiang, Schwartz L H, Zhao Binsheng. Segmentation of lung lesions on CT scans using watershed, active contours, and Markov random field [J].



Medical Physics, 2013, 40 (4): 043502-043502.

[7] El-Baz A, Nitzken M, Elnakib A, *et al.* 3D shape analysis for early diagnosis of malignant lung nodules. [C]// Proc of International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer-Verlag, 2011: 175-82.

[8] Breiman L. Random forests [J]. Machine learning, 2001, 45 (1): 5-32.

[9] Han Fangfang, Wang Huafeng, Zhang Guopeng, *et al.* Texture feature analysis for computer-aided diagnosis on pulmonary nodules [J]. Journal of Digital Imaging, 2015, 28 (1): 99-115.

[10] Dhara A K, Mukhopadhyay S, Dutta A, *et al.* A combination of shape and texture features for classification of pulmonary nodules in lung CT Images [J]. Journal of Digital Imaging, 2016, 29 (4): 466-475.

[11] Reeves A P, Xie Yiting, Jirapatnakul A. Automated pulmonary nodule CT image characterization in lung cancer screening [J]. International Journal of Computer Assisted Radiology & Surgery, 2016, 11 (1): 73-88.

[12] Maldonado F, Boland J M, Raghunath S, *et al.* Noninvasive characterization of the histopathologic features of pulmonary nodules of the lung adenocarcinoma spectrum using computer-aided nodule assessment and risk yield (CANARY) -pilot study [J]. Journal of Thoracic Oncology, 2013, 8 (4): 452-460.

[13] Rousseeuw P J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis [J]. Journal of Computational & Applied Mathematics, 1987, 20: 53-65.

[14] Sculley D. Web-scale k-means clustering [C]// Proc of International Conference on World Wide Web. 2010: 1177-1178.

[15] 裴博，强彦，赵涓涓. 一个基于 PET//CT 的孤立性肺结节恶性概率的预测模型 [J]. 计算机应用与软件, 2015, 32 (12): 170-174. (Pei Bo, Qiang Yan, Zhao Juanjuan. A PET//CT-Based prediction model for malignancy probability of solitary pulmonary nodules [J]. Computer Applications and Software, 2015, 32 (12): 170-174.)